Systematic and Applied Microbiology 37 (2014) 95–99

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Genomic insights into the taxonomic status of the three subspecies of *Bacillus subtilis*☆

Hana Yi[a,b], Jongsik Chun[c], Chang-Jun Cha[d,*]

[a] Department of Public Health Science, Graduate School, Korea University, Seoul, Republic of Korea
[b] Korea University Guro Hospital, Korea University, Seoul, Republic of Korea
[c] School of Biological Sciences and Bioinformatics Institute, BIO-MAX, Seoul National University, Seoul, Republic of Korea
[d] Department of Systems Biotechnology, Chung-Ang University, Anseong, Republic of Korea

ARTICLE INFO

Article history:
Received 15 March 2013
Received in revised form 14 August 2013
Accepted 27 September 2013

Keywords:
Bacillus subtilis
Comparative genomics
ANI
Subtilin
Bacillus subtilis subsp. *inaquosorum*

ABSTRACT

Bacillus subtilis contains three subspecies, i.e., subspecies *subtilis*, *spizizenii*, and *inaquosorum*. As these subspecies are phenotypically indistinguishable, their differentiation has relied on phylogenetic analysis of multiple protein-coding gene sequences. *B. subtilis* subsp. *inaquosorum* is a recently proposed taxon that encompasses strain KCTC 13429^T and related strains, which were previously classified as members of subspecies *spizizenii*. However, DNA–DNA hybridization (DDH) values among the three subspecies raised a question as to their independence. Thus, we evaluated the taxonomic status of subspecies *inaquosorum* using genome-based comparative analysis. In contrast to the previous experimental values of DDH, the inter-genomic relatedness inferred by average nucleotide identity (ANI) values indicated that subspecies *inaquosorum* and *spizizenii* were sufficiently different from subspecies *subtilis* and hence raised the possibility that the former two could be classified as separate species from *B. subtilis*. The genome-based tree also supported the separation of the two subspecies from *B. subtilis*. The exclusive presence of a subtilin synthesis system in subspecies *spizizenii* was a remarkable genetic characteristic that could even distinguish subspecies *spizizenii* from subspecies *inaquosorum* in addition to the low ANI values (<95%). Conclusively, the genome-based data obtained in this study demonstrated that subspecies *inaquosorum* and *spizizenii* are clearly distinguished from subspecies *subtilis*, and raises the possibility that these two subspecies could be classified as separate species from *B. subtilis*. In addition, the low ANI values between subspecies *inaquosorum* and *spizizenii* and the exclusive presence of subtilin synthesis genes in subspecies *spizizenii* also suggest circumscription of these two subspecies at the species level.

Introduction

Bacillus subtilis is divided into three subspecies, i.e., *Bacillus. subtilis* subsp. *subtilis* [17], *B. subtilis* subsp. *spizizenii* [17], and *B. subtilis* subsp. *inaquosorum* [22]. These three subspecies are phenotypically similar, and 16S rRNA gene sequences also fail to differentiate these subspecies due to the highly conserved nature of the gene [22]. Thus, the differentiation of these three subspecies has relied on phylogenetic analysis of multiple protein-coding gene sequences [21,22,26].

The presence of two distinct sub-groups within *B. subtilis* was firstly suggested by Roberts and Cohan in 1995 based on multi-gene phylogeny [21]. As these two groups showed DNA–DNA hybridization (DDH) values of 58–69% and sexual isolation, two novel subspecies were proposed by Nakamura et al. [17], i.e., *B. subtilis* subsp. *subtilis* and *B. subtilis* subsp. *spizizenii*. Further subspecies distinction was proposed for subspecies *spizizenii* by Rooney et al. (2009) based on multi-gene phylogeny [22]. The name *B. subtilis* subsp. *spizizenii* was used to accommodate the *B. subtilis* subsp. *spizizenii* sensu stricto strains. The name *B. subtilis* subsp. *inaquosorum* was proposed to accommodate the distinctive phylogenetic group encompassing strains B-23052^T (=KCTC 13429^T = NRRL B-23052^T) and B-14697 (=NRRL B-14697), which were previously classified as members of subspecies *spizizenii* [17]. According to Nakamura et al. [17], the DDH value between the type strains of subspecies *subtilis* and *spizizenii* was 63% and the DDH values between the two strains of subspecies *subtilis* and the two strains of subspecies *inaquosorum* were 65–68%. In contrast, subspecies *inaquosorum* strains B-23052^T and B-14697 showed a high DDH value of 90% with the type strain of subspecies *spizizenii*. Even though DDH values cannot be used for subspecies level distinction, the non-equivalent genomic relatedness among the three subspecies groups raised questions as to their distinctiveness at the subspecies levels.

☆ The GenBank accession number for the genome sequence of strain KCTC 13429^T is AMXN00000000.

* Corresponding author. Tel.: +82 31 670 4840; fax: +82 31 675 0432.
E-mail address: cjcha@cau.ac.kr (C.-J. Cha).

0723-2020/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.syapm.2013.09.006

In addition, subspecies *inaquosorum* was proposed mainly on the basis of five protein-coding gene sequences (*gyrA*, *rpoB*, *purH*, *polC*, and *groEL*) without genome-level comparison, e.g., DDH [22]. Therefore, the taxonomic status at the subspecies level should be re-evaluated using a genome-based method. As phenotypic differences have been found to be insignificant among many members of *B. subtilis*, differences in the genetic contents of the three subspecies are also of interest. However, the lack of genome sequences of subspecies *inaquosorum* has hindered the comparative study of these subspecies at the genomic level.

Thus, in this study, we determined the genome sequence of subspecies *inaquosorum* strain KCTC 13429^T, and revisited the taxonomic status of subspecies *inaquosorum* in the genomic era to see if this group is eligible for independent subspecies status. Our study also aimed to provide insight into the genetic characteristics of the three *B. subtilis* subspecies by conducting comparative genomic analysis.

Methods

Bacterial strain and DNA extraction

For genome sequencing, the type strain of *B. subtilis* subsp. *inaquosorum* was obtained from the Korean Collection for Type Cultures (KCTC 13429^T). Genomic DNA was extracted using a Wizard Genomic DNA Purification Kit (Promega).

Genome sequencing and annotation

The draft genome sequence was determined using a combination of paired-end shotgun sequencing (read length = 150 bp) on an Illumina Genome Analyzer IIx (3,118,224 reads; 108 × coverage; >Q30 = 91.9%) and single molecule sequencing (insert size = 5 Kb) on a Pacific Biosciences PacBio RS DNA sequencing system (75,380 reads; 33 × coverage; >Q30 = 83.0%). The sequencing reads were assembled using CLC genomics wb5.5 (CLCbio) and CodonCode Aligner (CodonCode Co.). The assembled contig sequences were uploaded into the RAST server [4] to predict the open reading frames (ORFs) by using Glimmer 3 [7]. The predicted ORFs were annotated by searching against clusters of orthologous group [24] and SEED databases [8].

Comparative genomics

For genome comparison, reference genome sequences belonging to *B. subtilis* were obtained from EzGenome Database (http://ezgenome.ezbiocloud.net). Comparative genomic analysis were conducted as described previously [6]. In brief, a segment of a target contig, which was homologous to a query ORF, was identified using the BLASTN program. This potentially homologous region was expanded by 2000 bp in both directions. Nucleotide sequences of the query ORF and target homologous regions were then aligned using a pairwise global alignment algorithm [16], and the resulting matched region in the subject contig was extracted and saved as a homolog. Orthologs and paralogs were differentiated by reciprocal comparison.

Calculation of ANI and construction of ANI dendrogram

The inter-genomic distances between genome sequences were determined from fully or partially sequenced genomes using average nucleotide identity (ANI) [13,20]. In a given pair of genomes, the query genome was cut into small pieces *in silico* (1020 bp). Then, high-scoring segment pairs between two genome sequences were determined using the BLAST algorithm [1,11]. The ANI was then calculated from the sets of high-scoring segment pairs. To convert the ANI into a distance, its complement to 1 was taken. From this pairwise distance matrix, an ANI dendrogram was constructed using the Unweighted Pair Group Method with Arithmetic Mean clustering method.

Construction of orthologous gene tree

A set of highly conserved orthologous ORFs (624 genes in total, 370,999 bp) showing greater than 95% nucleotide sequence similarity to *B. subtilis* subsp. *subtilis* NCIB3610^T (PRJNA55265) were selected as highly conserved protein-coding genes of the species *B. subtilis* and were then aligned using CLUSTALW [25]. The resultant multiple alignments were concatenated and then used to construct a genome tree with the neighbor-joining method [23] implemented in the MEGA program [14]. An evolutionary distance matrix for the neighbor-joining tree was generated according to the model of Jukes and Cantor [12]. The tree topologies were evaluated by bootstrap analysis [10] based on 1000 resamplings.

Analysis of codon usage bias

The calculation of the relative synonymous codon usage (RSCU) value and multivariate statistical analysis were performed using the CAIcal server [19] and R package 2.11.0 (www.r-project.org), respectively. Briefly, RSCU was determined by dividing the observed codon counts by the expected counts, where the expected counts assumed random usage of the synonymous codons for each amino acid. The determined RSCU values of codons among the genomic contigs or genes were ordinated using correspondence analysis implemented in R package.

Results and discussion

Genome sequence

The genome assembly of strain KCTC 13429^T generated 24 contigs within 23 scaffolds. The average Phred quality score of the final assembly was Q40. The genome sequence was deposited at GenBank under the accession number AMXN00000000. The genome size was 4.34 Mb with a G + C content of 43.69 mol%, and the genome contained 4439 predicted protein-coding sequences. The overall genomic characteristics of strain KCTC 13429^T were within the range of the genomes of *B. subtilis* (Table S1).

Genomic relatedness among the three subspecies groups

Determination of the ANI value between a given pair of genomes has been recognized as a robust means to reflect the degree of evolutionary distance between the compared genomes [11,13]. More recently, Richer and Rosselló-Móra [20] proposed that ANI could be used to substitute the 70% DDH standard of the current prokaryotic species definition, with a boundary that could be set at ~95–96%. In the present study, the ANI values among the three type strains of *B. subtilis* subspecies were all below the cut-off value for species circumscription (Table 1), suggesting that the three subspecies could belong to different genomic species. As we mentioned earlier, these results are in agreement with the previous DDH data (63–68%) between subspecies *subtilis* and two other subspecies [17]. However, a DDH value of 90% between subspecies *inaquosorum* strain KCTC 13429^T and the type strain of subspecies *spizizenii* may conflict with the ANI value (94.26–94.32%) calculated in this study (Table 1). When the ANI analysis were expanded to all available genomes that belong to the three subspecies groups, intra-subspecies genomic relatedness based on ANI was above the 95% boundary suggested by Richer and Rosselló-Móra [20], but the